Rachunek prawdopodobieństwa i statystyka

Wnioskowanie statystyczne

Maciej Smołka

Wydział Infomatyki AGH

Semestr zimowy 2024/25

Rodzaje zbieżności zmiennych losowych I

Niech $X_1, X_2, \ldots$ i X będą zmiennymi losowymi o wartościach w tej samej przestrzeni $\mathcal{X}$.

Definicja

Ciąg X_n jest **zbieżny do X według prawdopodobieństwa** (lub **stochastycznie**) jeśli dla każdego $\varepsilon > 0$

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1.$$

Definicja

Ciąg X_n jest **zbieżny do X z prawdopodobieństwem** 1 jeśli

$$P(\lim_{n\to\infty} X_n = X) = 1.$$

Rodzaje zbieżności zmiennych losowych II

Definicja

Załóżmy, że $\mathcal{X} = \mathbb{R}^k$ i niech F_{X_n} będzie dystrybuantą X_n, a F_X dystrybuantą X. Ciąg X_n jest **zbieżny do X według dystrybuant** jeśli dla każdego $x \in \mathbb{R}^k$ takiego, że F_X jest ciągła w x zachodzi warunek

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x).$$

Zależności między rodzajami zbieżności

1. Jeśli $X_n \to X$ z prawdopodobieństwem 1, to $X_n \to X$ według prawdopodobieństwa.
2. Jeśli $X_n \to X$ według prawdopodobieństwa, to istnieje podciąg X_{n_k} taki, że $X_{n_k} \to X$ z prawdopodobieństwem 1.
3. Jeśli $X_n \to X$ według prawdopodobieństwa, to $X_n \to X$ według dystrybuant.
4. Niech $c \in \mathbb{R}$ będzie stałą. $X_n \to c$ według prawdopodobieństwa wtedy i tylko wtedy, gdy $X_n \to c$ według dystrybuant.

Podstawowe zagadnienia wnioskowania statystycznego

Niech zmienna losowa X określa model rozkładu pewnej cechy (cech) w ustalonym zbiorze instancji (tzw. **populacji generalnej**). Innymi słowy, przyjmujemy, że wartości cech zachowują się jakby zostały wybrane *losowo* zgodnie z rozkładem zmiennej X. Do podstawowych zagadnień wnioskowania statystycznego należą:

- oszacowanie wielkości charakteryzujących rozkład X (np. wartości średniej albo wariancji);
- weryfikacja hipotez dotyczących rozkładu X (np. zgodności tego rozkładu z rozkładem normalnym).

Model statystyczny

Definicja

Modelem statystycznym nazywamy parę $(\mathcal{X}, \mathcal{P})$, gdzie $\mathcal{P}$ jest rodziną rozkładów prawdopodobieństwa (miar probabilistycznych) na zbiorze $\mathcal{X}$. Zazwyczaj przyjmuje się, że

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

dla pewnego zbioru **parametrów modelu** Θ. Mówimy, że model jest **identyfikowalny** jeśli $P_{\theta_1} \neq P_{\theta_2}$ dla $\theta_1 \neq \theta_2$.

Przykład

Załóżmy, że badamy zależność między wzrostem a wiekiem dzieci w wieku szkolnym, przy czym dysponujemy równomiernie wybraną próbką takich dzieci. Model rozkładu tych cech w badanej populacji stanowi zmienna losowa $(Wzrost, Wiek)$, przy czym wzrost jest liniowo zależny od wieku z uwzględnieniem błędu losowego o rozkładzie normalnym ze średnią 0, tzn.

$$Wzrost = b_0 + b_1 \cdot Wiek + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma)$$

dla pewnych b_0, b_1, σ. W tej sytuacji

$$\theta = (b_0, b_1, \sigma) \in \Theta \subset \mathbb{R}^3.$$

Prawdziwość modelu a jego użyteczność

Uwaga

Nie zakładamy, że rozkład zmiennej X należy do rodziny $\mathcal{P}$. W praktyce taka sytuacja zdarza się rzadko. Wobec tego trudno mówić o *prawdziwości* modelu statystycznego. Natomiast model powinien stanowić *użyteczne* uproszczenie lub aproksymację rzeczywistości.

Klasy modeli statystycznych

Model statystyczny nazywamy:

- **parametrycznym** jeśli $\Theta \subset \mathbb{R}^k$;
- **nieparametrycznym** jeśli $\Theta \subset V$ dla pewnej nieskończenie wymiarowej przestrzeni wektorowej V;
- **semi-parametrycznym** jeśli $\Theta \subset \mathbb{R}^k \times V$.

Prosta próba losowa

Definicja

Prostą próbą losową o liczności n nazywamy ciąg niezależnych zmiennych losowych $X_1, \ldots, X_n$ o wartościach w $\mathcal{X}$ i o tym samym rozkładzie $P_\theta \in \mathcal{P}$.

Definicja

Ciąg wartości $x_1, \ldots, x_n \in \mathcal{X}$ takich, że

$$X_1(\omega) = x_1, \ldots, X_n(\omega) = x_n$$

dla pewnego ω nazywamy **realizacją prostej próby losowej** $X_1, \ldots, X_n$.

Dystrybuanta empiryczna

W dalszym ciągu będziemy zakładać, że $\mathcal{X} = \mathbb{R}$.

Definicja

Dystrybuantą empiryczną nazywa się funkcję określoną wzorem

$$\hat{F}_n(x) = \frac{1}{n} \# \{ i \in \{1, \dots, n\} : x_i \leqslant x \}$$

gdzie $x_1, \dots, x_n$ jest realizacją prostej próby losowej.

Uwaga

Dla każdego $x \in \mathbb{R}$ $F_n(x)$ jest zmienną losową, tj.

$$\hat{F}_n(x, \omega) = \frac{1}{n} \# \{ i \in \{1, \dots, n\} : X_i(\omega) \leqslant x \}.$$

Asymptotyczne zachowanie dystrybuanty empirycznej

Następujące twierdzenie zostało sformułowane przez W. I. Gliwenkę[1] i F. P. Cantelli'ego [2].

Twierdzenie Gliwenki-Cantelli'ego

Jeśli F jest dystrybuantą rozkładu P_θ, to

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \to 0 \quad (n \to \infty)$$

z prawdopodobieństwem 1.

[1] **Walery Iwanowicz Gliwenko** (1897–1940) — ukraiński matematyk

[2] **Francesco Paolo Cantelli** (1875–1966) — włoski matematyk

Statystyki

Definicja

Statystyką nazywa się zmienną losową będącą funkcją prostej próby losowej $T(X_1, \ldots, X_n)$.
Średnią w prostej próbie losowej nazywa się statystykę

$$\overline{X} = \overline{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

Wariancją w prostej próbie losowej nazywa się statystykę

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2.$$

Prawo wielkich liczb

Jeśli P_θ ma skończoną wartość średnią μ i wariancję, to

$$\overline{X}_n \to \mu$$

z prawdopodobieństwem 1 przy $n \to \infty$.

Wielkości opisujące rozkład średniej

Niech rozkład P_θ ma wartość średnią μ i wariancję σ^2. Ponieważ $X_1, \ldots, X_n$ są niezależne i mają rozkład P_θ, więc

$$E(\overline{X}) = \frac{1}{n} EX_1 + \cdots + \frac{1}{n} EX_n = \mu$$
$$Var(\overline{X}) = \frac{1}{n^2} VarX_1 + \cdots + \frac{1}{n^2} VarX_n = \frac{\sigma^2}{n}$$
$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Centralne Twierdzenie Graniczne

Oznaczmy

$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

Następujące twierdzenie nosi również nazwę twierdzenia Lindeberga[3]-Lévy'ego[4].

Twierdzenie Lindeberga-Lévy'ego

Niech X_n będzie ciągiem niezależnych zmiennych losowych o tym samym rozkładzie ze skończoną średnią μ i wariancją $\sigma^2 > 0$. Wtedy ciąg Z_n jest zbieżny według dystrybuant do zmiennej losowej Z o standardowym rozkładzie normalnym $\mathcal{N}(0, 1)$.

[3] **Jarl Waldemar Lindeberg** (1876–1932) — fiński matematyk
[4] **Paul Pierre Lévy** (1886–1971) — francuski matematyk

Wniosek dla prostej próby losowej

Jeśli $\Phi = \Phi_{0,1}$ jest dystrybuantą standardowego rozkładu normalnego, to dla dowolnych $a, b \in \mathbb{R}$ takich, że $a \leqslant b$ zachodzi

$$P\left(a \leqslant \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leqslant b\right) \to P(a \leqslant Z \leqslant b) = \Phi(b) - \Phi(a) \quad (n \to \infty).$$

Inaczej mówiąc, prawdopodobieństwo $P(a \leqslant \overline{X} \leqslant b)$ jest dla dużych liczności prób w przybliżeniu równe

$$P\left(\frac{a-\mu}{\frac{\sigma}{\sqrt{n}}} \leqslant Z \leqslant \frac{b-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(a \leqslant \frac{\sigma}{\sqrt{n}} Z + \mu \leqslant b\right),$$

czyli rozkład średniej dla dużych n jest w przybliżeniu równy rozkładowi normalnemu $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$.

Uwaga

Przybliżenie normalne dla rozkładu średniej zasadniczo uważa się za wystarczająco dokładne dla $n \geqslant 25$.

Estymatory punktowe

Definicja

Estymatorem (lub **estymatorem punktowym**) parametru θ nazywa się statystykę $\hat{\theta}(X_1, \ldots, X_n)$ służącą do oszacowania wartości θ. Liczbę $\hat{\theta}(x_1, \ldots, x_n)$ dla konkretnej realizacji próby losowej nazywa się **wartością estymatora** albo **estymatą**.

Obciążenie estymatora

Liczbę

$$Bias(\hat{\theta}) = E\hat{\theta} - \theta = E\hat{\theta}(X_1, \ldots, X_n) - \theta$$

nazywamy **obciążeniem estymatora** [ang. *bias*]. Jeśli $Bias(\hat{\theta}) = 0$, to estymator nazywamy **nieobciążonym** [ang. *unbiased*], w przeciwnym razie estymator nazywamy **obciążonym** [ang. *biased*]. Estymator nazywamy **asymptotycznie nieobciążonym** jeśli

$$\lim_{n\to\infty} Bias(\hat{\theta}) = 0.$$

Estymator wartości średniej

Wiemy, że

$$E\overline{X} = \mu,$$

zatem $\overline{X}$ jest nieobciążonym estymatorem wartości średniej μ.

Estymatory wariancji I

Estymatorem σ^2 jest wariancja w próbie, tzn.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right).$$

Ponieważ $\sigma_X^2 = \mu_{X^2} - \mu_X^2$, więc $\mu_{X^2} = \sigma_X^2 + \mu_X^2$, a zatem oznaczając $Y = \sum_{i=1}^{n} X_i$ otrzymujemy

$$\begin{aligned} E(S^2) &= \frac{1}{n-1}\left(\sum_{i=1}^{n} EX_i^2 - \frac{1}{n}EY^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}(\sigma^2 + \mu^2) - \frac{1}{n}(\sigma_Y^2 + \mu_Y^2)\right) \\ &= \frac{1}{n-1}\left(n\sigma^2 + n\mu^2 - \frac{1}{n}n\sigma^2 - \frac{1}{n}n^2\mu^2\right) = \sigma^2, \end{aligned}$$

Estymatory wariancji II

co oznacza, że S^2 jest nieobciążonym estymatorem σ^2. Konsekwentnie,

$$M_2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{n-1}{n}S^2$$

jest obciążonym estymatorem wariancji, ponieważ

$$Bias(M_2) = E(M_2) - \sigma^2 = \frac{n-1}{n}E(S^2) - \sigma^2 = -\frac{1}{n}\sigma^2.$$

Jest to natomiast oczywiście estymator asymptotycznie nieobciążony.

Estymatory nieobciążone o minimalnej wariancji

Definicja

Estymatorem nieobciążonym o minimalnej wariancji [ang. *minimum-variance unbiased estimator (MVUE)*] nazywamy taki estymator nieobciążony $\hat{\theta}$, że dla dowolnej wartości parametru $\theta \in \Theta$ i dowolnego estymatora nieobciążonego $\tilde{\theta}$ zachodzi nierówność

$$Var(\hat{\theta}) \leqslant Var(\tilde{\theta}).$$

Przykłady

1. $\overline{X}$ jest MVUE w modelu

$$\mathcal{P} = \{\mathcal{N}(\mu, \sigma_0) : \mu \in \mathbb{R}\}.$$

dla ustalonego $\sigma_0 > 0$.

2. $\overline{X}$ i S^2 są MVUE dla μ i σ^2 w modelu

$$\mathcal{P} = \{\mathcal{N}(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}.$$

Błąd średniokwadratowy

Definicja

Błędem średniokwadratowym [ang. *mean squared error (MSE)*] estymatora $\hat{\theta}$ parametru θ nazywamy liczbę

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2.$$

Estymatory dominujące i dopuszczalne

Definicja

Mówimy, że **estymator** θ_1 **dominuje nad** θ_2 jeśli dla każdej wartości $\theta \in \Theta$

$$MSE(\theta_1) \leqslant MSE(\theta_2)$$

i istnieje taka wartość θ, dla której nierówność jest ostra.
Estymator nazywamy **dopuszczalnym** jeśli nie istnieje estymator nad nim dominujący.

Związek między obciążeniem a MSE

Stwierdzenie

Dla dowolnego estymatora $\hat{\theta}$ zachodzi równość

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias(\hat{\theta})^2.$$

Dowód stwierdzenia

Oznaczmy $\mu_{\hat{\theta}} = E(\hat{\theta})$. Wówczas mamy

$$(\hat{\theta} - \theta)^2 = (\hat{\theta} - \mu_{\hat{\theta}} + \mu_{\hat{\theta}} - \theta)^2 = (\hat{\theta} - \mu_{\hat{\theta}})^2 + 2(\hat{\theta} - \mu_{\hat{\theta}})(\mu_{\hat{\theta}} - \theta) + (\mu_{\hat{\theta}} - \theta)^2.$$

Licząc wartość średnią obu stron otrzymujemy

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + 2(\mu_{\hat{\theta}} - \theta)E(\hat{\theta} - m_{\hat{\theta}}) + Bias(\hat{\theta})^2 = Var(\hat{\theta}) + Bias(\hat{\theta})^2.$$

Wnioski

1. MVUE ma najmniejszy MSE spośród estymatorów nieobciążonych danego parametru.
2. W danym modelu może istnieć estymator obciążony o mniejszym MSE od dowolnego estymatora nieobciążonego.

Przykład

W rodzinie rozkładów normalnych S^2 jest nieobciążonym estymatorem wariancji i

$$MSE(S^2) = Var(S^2) = \frac{2\sigma^4}{n-1}$$

Natomiast dla innego estymatora wariancji $T^2 = \frac{n-1}{n+1} S^2$ mamy

$$MSE(T^2) = \frac{2\sigma^4}{n+1} < MSE(S^2),$$

pomimo że estymator jest obciążony.

Błąd standardowy

Definicja

Błędem standardowym $SE(\hat{\theta})$ estymatora $\hat{\theta}$ nazywamy albo jego odchylenie standardowe, albo dowolny estymator tego odchylenia standardowego.

Przykład

$SE(\overline{X}) = \frac{S}{\sqrt{n}}$, gdzie $S = \sqrt{S^2}$.

Funkcja wiarygodności

Definicja

Funkcją wiarygodności [ang. *likelihood function*] dla modelu $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ nazywamy funkcję

$$\mathcal{L} : \Theta \times \mathbb{R}^n \ni (\theta, x) \longmapsto \mathcal{L}(\theta|x) \in [0, +\infty)$$

wyznaczającą rozkład łączny obserwowanych danych jako funkcję parametru θ.

Funkcja wiarygodności dla prostej próby losowej

Niech $X_1, \ldots, X_n$ będzie prostą próbą losową. Jeśli P_θ są rozkładami ciągłymi o gęstościach f_θ, to

$$\mathcal{L}(\theta|x) = f_\theta(x_1) \cdots f_\theta(x_n).$$

Jeśli P_θ są rozkładami dyskretnymi z funkcjami prawdopodobieństwa p_θ, to

$$\mathcal{L}(\theta|x) = p_\theta(x_1) \cdots p_\theta(x_n).$$

Logarytmiczna funkcja wiarygodności

Ze względów obliczeniowych często rozważa się *logarytmiczną funkcję wiarygodności* [ang. *log-likelihood function*], tzn.

$$l(\theta|x) = \ln \mathcal{L}(\theta|x).$$

Dla rozkładów ciągłych mamy

$$l(\theta|x) = \ln f_\theta(x_1) + \cdots + \ln f_\theta(x_n),$$

a dla rozkładów dyskretnych

$$l(\theta|x) = \ln p_\theta(x_1) + \cdots + \ln p_\theta(x_n),$$

Przykład

Niech

$$\mathcal{P} = \{\mathcal{N}(\theta_1, \sqrt{\theta_2}) : \theta_1 \in \mathbb{R}, \theta_2 > 0\}.$$

Wtedy mamy

$$\mathcal{L}(\theta_1, \theta_2 | x) = \frac{1}{\sqrt{(2\pi\theta_2)^n}} e^{-\frac{1}{2\theta_2}\sum_{i=1}^n (x_i - \theta_1)^2}$$

oraz

$$l(\theta_1, \theta_2 | x) = -\frac{n}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}\sum_{i=1}^n (x_i - \theta_1)^2.$$

Przykład

Dla

$$\mathcal{P} = \{Pois(\theta) : \theta > 0\}$$

mamy

$$\mathcal{L}(\theta|x) = e^{-n\theta}\frac{\theta^{x_1+\cdots+x_n}}{x_1!\cdots x_n!}$$

oraz

$$l(\theta|x) = -n\theta + \left(\sum_{i=1}^{n} x_i\right)\ln\theta - \sum_{i=1}^{n}\ln(x_i!).$$

Metoda największej wiarygodności

Jest to metoda konstrukcji estymatorów opracowana przez R. Fishera[5]. **Estymatorem największej wiarygodności** [ang. *maximum likelihood estimator (MLE)*] nazywamy funkcję $\hat{\theta}$, która przy ustalonych wartościach obserwacji $x = (x_1, \ldots, x_n)$ maksymalizuje wartość funkcji wiarygodności lub, alternatywnie, wartość logarytmicznej funkcji wiarygodności, tzn.

$$\mathcal{L}(\hat{\theta}(x)|x) = \max_{\theta \in \Theta} \mathcal{L}(\theta|x) \quad \text{albo} \quad l(\hat{\theta}(x)|x) = \max_{\theta \in \Theta} l(\theta|x).$$

[5] **Ronald Aylmer Fisher** (1890–1962) — brytyjski statystyk i genetyk

Obliczanie MLE

Jeśli funkcja wiarygodności jest różniczkowalna względem $\theta = (\theta_1, \dots, \theta_k) \in \mathbb{R}^k$ dla dowolnego x, to MLE można czasem wyznaczyć analitycznie korzystając z warunku koniecznego optymalności, tzn. rozwiązując układ równań

$$\frac{\partial}{\partial \theta_1} \mathcal{L}(\theta|x) = \cdots = \frac{\partial}{\partial \theta_k} \mathcal{L}(\theta|x) = 0$$

albo

$$\frac{\partial}{\partial \theta_1} l(\theta|x) = \cdots = \frac{\partial}{\partial \theta_k} l(\theta|x) = 0$$

Jeśli MLE nie da się wyliczyć analitycznie, wyznacza się je przy użyciu algorytmów optymalizacji.

Przykład

Dla $\mathcal{P} = \{Pois(\theta) : \theta > 0\}$ mamy

$$\frac{\partial}{\partial\theta} l(\theta|x) = -n + \left(\sum_{i=1}^{n} x_i\right)\frac{1}{\theta}.$$

Zatem MLE spełnia równanie

$$-n + \left(\sum_{i=1}^{n} x_i\right)\frac{1}{\hat{\theta}} = 0,$$

skąd otrzymujemy

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}.$$

Własności MLE

1. MLE są asymptotycznie nieobciążone.
2. Jeśli $\hat{\theta}$ jest MLE parametru θ i h jest odwzorowaniem wzajemnie jednoznacznym, to $h(\hat{\theta})$ jest MLE parametru $h(\theta)$.

Metoda momentów

Załóżmy, że dla każdego $\theta = (\theta_1, \ldots, \theta_k)$ rozkłady P_θ mają skończone momenty do rzędu k oraz że istnieją funkcje h_p takie, że

$$EX_i^p = h_p(\theta_1, \ldots, \theta_k)$$

dla $p = 1, \ldots, k$. **Estymatory metody momentów** opracowanej przez K. Pearsona[6] otrzymujemy rozwiązując równania

$$\frac{1}{n}\sum_{i=1}^{n} X_i^p = h_p(\hat{\theta}_1, \ldots, \hat{\theta}_k).$$

[6] **Karl (Carl) Pearson** (1857–1936) — angielski matematyk i biostatystyk

Uwagi

1. Momenty zwykłe można zastąpić momentami centralnymi. Wtedy równania metody momentów mają postać
$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^p = \tilde{h}_p(\hat{\theta}_1, \dots, \hat{\theta}_k).$$
2. Drugi moment centralny M_2 można zastąpić wariancją w próbie S^2.
3. Estymatory metody momentów na ogół nie mają zbyt dobrych własności, ale są stosunkowo łatwe do wyznaczenia. Mogą być stosowane np. jako punkty startowe dla algorytmów optymalizacji w metodzie największej wiarygodności.

Przykład

Niech

$$\mathcal{P} = \{Gamma\,(\theta_1, \theta_2) : \theta_1 > 0, \theta_2 > 0)\}.$$

Wtedy

$$E(X_i) = \frac{\theta_1}{\theta_2} \quad Var(X_i) = \frac{\theta_1}{\theta_2^2}.$$

Biorąc S^2 zamiast M_2 otrzymujemy równania metody momentów

$$\overline{X} = \frac{\hat{\theta}_1}{\hat{\theta}_2} \quad S^2 = \frac{\hat{\theta}_1}{\hat{\theta}_2^2}.$$

A stąd

$$\hat{\theta}_1 = \frac{(\overline{X})^2}{S^2} \quad \hat{\theta}_2 = \frac{\overline{X}}{S^2}.$$

Przedział ufności dla rozkładu ciągłego

Pojęcie przedziału ufności zostało wprowadzone przez J. Neymana[7].

Definicja

Załóżmy, że rozkłady P_θ są ciągłe. Niech $\underline{U}, \overline{U}$ będą statystykami takimi, że $P(\underline{U} \leqslant \overline{U}) = 1$ i $0 < \alpha < 1$. **Przedziałem ufności** [ang. *confidence interval*] dla parametru θ na poziomie ufności [ang. *confidence level*] $1 - \alpha$ nazywamy przedział losowy $\left[\underline{U}, \overline{U}\right]$ o ile spełniona jest równość

$$P\left(\underline{U} \leqslant \theta \leqslant \overline{U}\right) = 1 - \alpha$$

[7] **Jerzy Spława-Neyman** (1894–1981) — polski matematyk i statystyk

Przedział ufności dla rozkładu dykretnego

Jeśli rozkłady P_θ są dyskretne, to modyfikuje się definicję przedziału ufności zmieniając równość w ostatnim warunku na nierówność, tzn. wymagamy tylko aby

$$P\left(\underline{U} \leqslant \theta \leqslant \overline{U}\right) \geqslant 1 - \alpha$$

Uwagi

1. Przedział ufności nie jest stałym przedziałem, tylko zmienną losową.
2. Standardowo stosowany poziom ufności to 95% (inaczej 0,95, czyli $\alpha = 0{,}05$), inne często stosowane to 99% i 90%.

Konstrukcja przedziału ufności

Wybieramy estymator punktowy U parametru θ, którego rozkład dokładny lub asymptotyczny jest znany. Następnie bierzemy liczby a i b takie, że

$$P(a \leqslant U \leqslant b) = 1 - \alpha.$$

Jeśli nierówność $a \leqslant U \leqslant b$ da się przekształcić do równoważnej nierówności

$$\underline{f}(U) \leqslant \theta \leqslant \overline{f}(U),$$

to $\left[\underline{f}(U), \overline{f}(U)\right]$ jest oczekiwanym przedziałem ufności na poziomie ufności $1 - \alpha$.

Uwaga

Zaprezentowana konstrukcja nie daje jednoznacznego określenia przedziału ufności, bo liczby a i b można dobrać na wiele sposobów. Zwykle, jeśli nie ma innych istotnych przesłanek, dobiera się je tak, aby

$$P(U < a) = P(U > b) = \frac{1}{2}\alpha$$

W przypadku rozkładów dyskretnych ten warunek zamienia się na

$$P(U < a) \leqslant \frac{1}{2}\alpha, \quad P(U > b) \leqslant \frac{1}{2}\alpha.$$

Średnia w rozkładzie normalnym o znanym σ I

Niech $\mathcal{P} = \{\mathcal{N}(\mu, \sigma) : \mu \in \mathbb{R}\}$ dla pewnego $\sigma > 0$. Szukamy przedziału ufności dla μ. MVUE (i MLE) w tym przypadku jest $\overline{X}$. Wiemy, że $\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$, a zatem

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1).$$

Jeśli oznaczymy przez z_p kwantyl rzędu p rozkładu $\mathcal{N}(0, 1)$, to otrzymamy

$$P(Z < z_{\frac{\alpha}{2}}) = \frac{\alpha}{2} \quad P(Z > z_{1-\frac{\alpha}{2}}) = 1 - P(Z \leqslant z_{1-\frac{\alpha}{2}}) = \frac{\alpha}{2}.$$

Stąd

$$P(z_{\frac{\alpha}{2}} \leqslant Z \leqslant z_{1-\frac{\alpha}{2}}) = 1 - \alpha.$$

Średnia w rozkładzie normalnym o znanym σ II

Z symetrii rozkładu $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$, czyli

$$P(-z_{1-\frac{\alpha}{2}} \leqslant Z \leqslant z_{1-\frac{\alpha}{2}}) = 1-\alpha.$$

Zatem

$$P(\overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1-\alpha.$$

Oznacza to, że przedziałem ufności na poziomie $1-\alpha$ dla μ jest

$$\left[\overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$$

Uwagi

1. Rezultatem podanej konstrukcji jest tzw. **dwustronny** przedział ufności. Zamiast tego czasem konstruuje się przedziały **jednostronne**, np.
$$\left(-\infty, \overline{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right] \quad \text{lub} \quad \left[\overline{X} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}, +\infty\right).$$
2. Spośród przedziałów dwustronnych o tym samym poziomie ufności przedział symetryczny z poprzedniego slajdu jest najkrótszy.

Średnia w rozkładzie normalnym o nieznanym σ

Tym razem $\mathcal{P} = \{\mathcal{N}(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}$. Do konstrukcji przedziału ufności zamiast zmiennej Z bierzemy

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}.$$

Można pokazać, że

$$T \sim t_{n-1},$$

gdzie t_n jest tzw. **rozkładem** t o n stopniach swobody.

Rozkład t

Rozkład t o ν stopniach swobody dla $\nu > 0$, zwany także **rozkładem Studenta**[8], to rozkład ciągły o gęstości

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

[8]**William Sealy Gosset** (1876–1937) – brytyjski statystyk, ze względu na ograniczenia stawiane przez pracodawcę publikacje naukowe podpisywał pseudonimem **Student**

Alternatywne określenie rozkładu t

Jeśli zmienne losowe Z i V są niezależne, a ponadto

$$Z \sim \mathcal{N}(0, 1)$$

i

$$V \sim \chi^2(\nu),$$

to

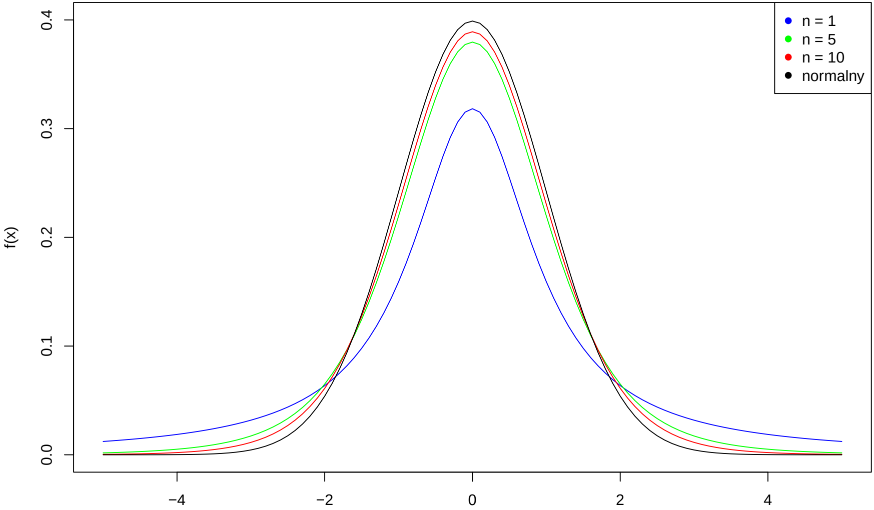
$$\frac{Z}{\sqrt{\frac{V}{\nu}}} = Z\sqrt{\frac{\nu}{V}} \sim t_\nu.$$

Wielkości opisujące rozkład t

1. Średnia dla $\nu \leqslant 1$ nie istnieje, dla $\nu > 1$ jest równa 0.
2. Wariancja jest równa $\frac{\nu}{\nu-2}$ dla $\nu > 2$, $+\infty$ dla $1 < \nu \leqslant 2$, nie istnieje dla $\nu \leqslant 1$.
3. Współczynnik asymetrii jest równy 0 dla $\nu > 3$, nie istnieje dla $\nu \leqslant 3$.
4. Współczynnik wyostrzenia przyjmuje następujące wartości

$$\gamma_2 = \begin{cases} \frac{6}{\nu-4} & \text{dla } \nu > 4, \\ +\infty & \text{dla } 2 < \nu \leqslant 4, \\ \text{nie istnieje} & \text{dla } \nu \leqslant 2. \end{cases}$$

Gęstość rozkładu t



Średnia w rozkładzie normalnym o nieznanym σ (c.d.)

Skoro wiemy, że $T \sim t_{n-1}$, więc analogicznie jak dla przypadku ze znanym σ korzystając z symetrii rozkładu t otrzymujemy symetryczny przedział ufności na poziomie $1-\alpha$ o następującej postaci

$$\left[\overline{X} - t_{1-\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \overline{X} + t_{1-\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right],$$

gdzie $t_{p,\nu}$ jest kwantylem rzędu p rozkładu t_ν.

Uwaga

Dla $n \to \infty$ rozkład t coraz bardziej przypomina standardowy rozkład normalny. Dla $n = 30$ te rozkłady są praktycznie nieodróżnialne.

Porównanie średnich w rozkładach normalnych — znane wariancje

Jeśli mamy dwie **niezależne próby losowe** $X_1, \ldots, X_{n_1}$ i $Y_1, \ldots, Y_{n_2}$ o nieznanych średnich m_1 i m_2, i znanych odchyleniach standardowych σ_1 i σ_2, to

$$Z = \frac{(\overline{X} - \overline{Y}) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1).$$

Stąd przedział ufności na poziomie $1 - \alpha$ jest równy

$$\left[\overline{X} - \overline{Y} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \overline{X} - \overline{Y} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right].$$

Porównanie średnich w rozkładach normalnych — nieznane wariancje I

Zakładamy, że $\sigma_1 = \sigma_2 = \sigma$. Wtedy

$$Var(X - Y) = \frac{\sigma^2}{n_1 + n_2}$$

i można wykazać, że

$$S_p^2 = \frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}$$

jest nieobciążonym estymatorem σ^2, przy czym

$$T = \frac{(\overline{X} - \overline{Y}) - (m_1 - m_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Porównanie średnich w rozkładach normalnych — nieznane wariancje II

Stąd przedział ufności na poziomie $1 - \alpha$ jest równy

$$\left[\overline{X} - \overline{Y} - t_{1-\frac{\alpha}{2}, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \right.$$
$$\left. \overline{X} - \overline{Y} + t_{1-\frac{\alpha}{2}, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right].$$

Porównanie średnich w rozkładach normalnych — pary obserwacji

Zakładamy, że pary $(X_1, Y_1), \ldots, (X_n, Y_n)$ mają takie same dwuwymiarowe rozkłady normalne i są wzajemnie niezależne, ale X_i i Y_i mogą być zależne. W takiej sytuacji zmienne

$$D_i = X_i - Y_i$$

są niezależne i mają rozkład normalny, zatem można jak w jednym z poprzednich przykładów wyznaczyć przedział ufności dla średniej m_D przy pomocy kwantyli standardowego rozkładu normalnego.

Wariancja w rozkładzie normalnym I

Jeśli $X_i \sim \mathcal{N}(0,1)$ dla każdego i, to

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n).$$

Podobnie jeśli $X_i \sim \mathcal{N}(\mu, \sigma)$, to

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n).$$

Ponadto można pokazać, że

$$\mathcal{X}^2 = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Wariancja w rozkładzie normalnym II

Stąd

$$P(\chi^2_{\frac{\alpha}{2},n-1} \leqslant \mathcal{X}^2 \leqslant \chi^2_{1-\frac{\alpha}{2},n-1}) = 1-\alpha,$$

gdzie $\chi^2_{p,n}$ jest kwantylem rzędu p rozkładu $\chi^2(n)$. Korzystając z definicji zmiennej $\mathcal{X}^2$ otrzymujemy

$$P\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \leqslant \sigma^2 \leqslant \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}\right) = 1-\alpha.$$

Stąd otrzymujemy przedział ufności dla wariancji

$$\left[S^2\frac{n-1}{\chi^2_{1-\frac{\alpha}{2},n-1}}, S^2\frac{n-1}{\chi^2_{\frac{\alpha}{2},n-1}}\right]$$

Wariancja w rozkładzie normalnym III

i dla odchylenia standardowego

$$\left[S\sqrt{\frac{n-1}{\chi^2_{1-\frac{\alpha}{2},n-1}}}, S\sqrt{\frac{n-1}{\chi^2_{\frac{\alpha}{2},n-1}}} \right].$$

Rozkład F

Jeśli $U \sim \chi^2(n_1)$ i $V \sim \chi^2(n_2)$, przy czym zmienne są niezależne, to

$$\frac{\frac{U}{n_1}}{\frac{V}{n_2}} \sim F(n_1, n_2),$$

gdzie $F(n_1, n_2)$ jest **rozkładem F** o n_1 i n_2 stopniach swobody, zwanym też **rozkładem Snedecora**[9] albo **rozkładem Fishera-Snedecora**.

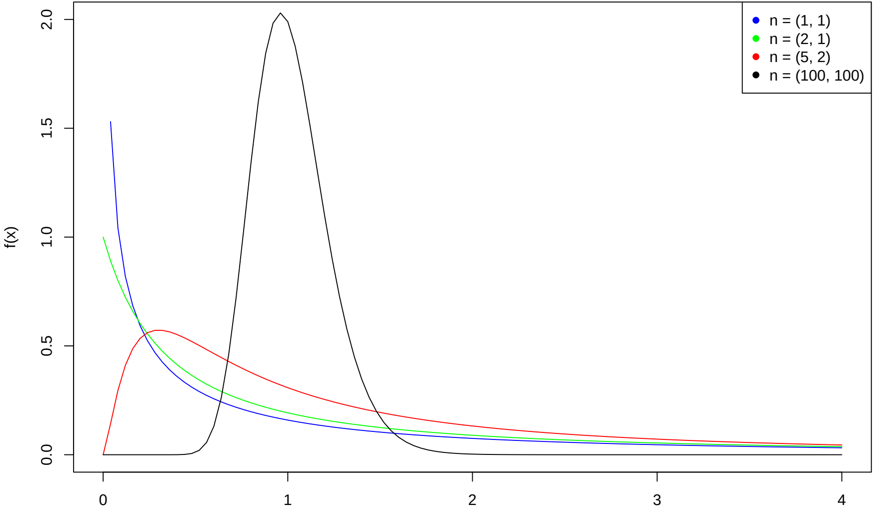
[9] **George Waddel Snedecor** (1881–1974) — amerykański matematyk i statystyk

Definicja rozkładu F

Jest to rozkład ciągły o gęstości

$$f(x) = \begin{cases} \dfrac{\Gamma\left(\frac{n_1+n_2}{2}\right)\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)\left(1+\frac{n_1}{n_2}x\right)^{\frac{n_1+n_2}{2}}} & \text{dla } x > 0, \\ 0 & \text{dla } x \leqslant 0. \end{cases}$$

Gęstość rozkładu F



Własności rozkładu F

1. Jeśli $X \sim F(n_1, n_2)$, to $\frac{1}{X} \sim F(n_2, n_1)$.
2. Jeśli $X \sim t_n$, to $X^2 \sim F(1, n)$.
3. Jeśli f_{p,n_1,n_2} jest kwantylem rzędu p rozkładu $F(n_1, n_2)$, to

$$f_{p,n_1,n_2} = \frac{1}{f_{1-p,n_2,n_1}}.$$

Porównanie wariancji rozkładów normalnych I

Zakładamy, że dane są dwie niezależne próby losowe o licznościach n_1 i n_2 z rozkładów $\mathcal{N}(m_1, \sigma_1)$ i $\mathcal{N}(m_2, \sigma_2)$. Ponieważ

$$\frac{(n_i - 1)S_i^2}{\sigma_i^2} \sim \chi^2(n_i - 1),$$

więc

$$F = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \sim F(n_1 - 1, n_2 - 1).$$

Stąd

$$P\left(f_{\frac{\alpha}{2}, n_1-1, n_2-1} \leqslant \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \leqslant f_{1-\frac{\alpha}{2}, n_1-1, n_2-1}\right) = 1 - \alpha.$$

Porównanie wariancji rozkładów normalnych II

Zatem przedział ufności dla ilorazu $\frac{\sigma_2^2}{\sigma_1^2}$ na poziomie ufności $1-\alpha$ ma postać

$$\left[\frac{S_2^2}{S_1^2} f_{\frac{\alpha}{2},n_1-1,n_2-1}, \frac{S_2^2}{S_1^2} f_{1-\frac{\alpha}{2},n_1-1,n_2-1}\right] =$$

$$\left[\frac{S_2^2}{S_1^2} \frac{1}{f_{1-\frac{\alpha}{2},n_2-1,n_1-1}}, \frac{S_2^2}{S_1^2} f_{1-\frac{\alpha}{2},n_1-1,n_2-1\cdot}\right]$$

Uwagi o innych rozkładach

- Z Centralnego Twierdzenia Granicznego wynika, że dla dużych n statystyka Z ma w przybliżeniu rozkład $\mathcal{N}(0, 1)$, zatem można na jej podstawie konstruować (przybliżone) przedziały ufności dla średniej w przypadku znanej wariancji. W praktyce za wystarczające uważa się $n \geqslant 30$, dla rozkładów wyraźnie skośnych $n \geqslant 40$.
- Analogicznie w przypadku nieznanej wariancji do konstrukcji przedziału ufności dla średniej stosuje się statystykę T przyjmując, że dla niezbyt małych wartości n ($n \approx 20$) ma ona rozkład zbliżony do rozkładu t.
- Przedstawiona metoda konstrukcji przedziałów ufności dla wariancji i ilorazu wariancji nie dopuszcza podobnego uogólnienia.

Hipotezy statystyczne

Rozważamy model statystyczny $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$.

Definicja

Hipotezą statystyczną nazywamy dowolny niepusty podzbiór $\mathcal{P}$.
W praktyce wyróżniamy jedną hipotezę zwaną **hipotezą zerową** (H_0), która podlega weryfikacji, natomiast $H_1 = \mathcal{P} \setminus H_0$ nazywamy **hipotezą alternatywną**.
Hipotezę jednoelementową nazywamy **hipotezą prostą**, wieloelementową nazywamy **hipotezą złożoną**.

Hipotezy dotyczące parametrów rozkładu

Jeśli hipotezy dotyczą nieznanej wartości parametru θ, to zbiór możliwych hipotez jest jednoznacznie wyznaczony przez zbiór możliwych wartości parametru. Można zatem utożsamić te dwa zbiory, tzn. przyjąć, że

$$H_0 \subset \Theta, \quad H_1 = \Theta \setminus H_0.$$

Weryfikacja hipotezy zerowej

Weryfikacja hipotezy zerowej polega na:

- skonstruowaniu statystyki U o znanym rozkładzie dokładnym lub asymptotycznym;
- ustaleniu zbioru C tych wartości U, których wystąpienie uznaje się za wspierające **odrzucenie hipotezy zerowej na korzyść hipotezy alternatywnej**.

Definicja

Statystykę U nazywamy **testem hipotezy zerowej** H_0 **przeciwko hipotezie alternatywnej** H_1, a zbiór C **zbiorem krytycznym testu**.

Błędy testów statystycznych

Definicja

Błędem pierwszego rodzaju nazywa się odrzucenie H_0, gdy jest ona prawdziwa. Prawdopodobieństwo wystąpienia błędu pierwszego rodzaju, tzn.

$$\alpha = P(U \in C | H_0)$$

nazywa się **poziomem istotności testu**.
Błędem drugiego rodzaju nazywa się odrzucenie H_1, gdy jest ona prawdziwa dla pewnej alternatywnej wartości θ. Dla zadanej alternatywnej wartości θ prawdopodobieństwo niepopełnienia błędu drugiego rodzaju nazywa się **mocą testu**, tzn.

$$1 - \beta = P(U \in C | H_1(\theta)).$$

Uwagi

1. Moc testu można zwiększać albo przez zwiększenie liczności próby, albo przez zwiększenie poziomu istotności.
2. Typowo stosowane poziomy istotności to: 0,05; 0,02; 0,01; 0,001.

Test dla wartości średniej przy znanej wariancji I

Mamy $\mathcal{P} = \{\mathcal{N}(\mu, \sigma) : \mu \in \mathbb{R}\}$ dla ustalonego $\sigma > 0$. Ustalamy hipotezy

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

Jako statystykę testową bierzemy

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}},$$

o której wiemy, że ma rozkład $\mathcal{N}(0, 1)$ o ile prawdziwa jest H_0. Jeśli H_1 jest prawdziwa dla pewnego μ, to

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} + \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}\left(\frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}}, 1\right).$$

Test dla wartości średniej przy znanej wariancji II

Zatem dla zadanego poziomu istotności α mamy

$$P(Z \leqslant -z_{1-\frac{\alpha}{2}} \text{ lub } Z \geqslant z_{1-\frac{\alpha}{2}} | H_0) = \alpha,$$

czyli możemy przyjąć

$$C = (-\infty, -z_{1-\frac{\alpha}{2}}] \cup [z_{1-\frac{\alpha}{2}}, +\infty).$$

Np. dla poziomu istotności 0,01 mamy

$$C = (-\infty, -2.5758293] \cup [2.5758293, +\infty).$$

p-wartości

Definicja

p-**wartością** dla zaobserwowanej wartości statystyki testowej nazywamy najmniejszy poziom istotności, przy którym ta wartość prowadzi do odrzucenia hipotezy zerowej.

Przykład

Dla testu równości wartości średnich przy znanej wariancji z dwustronną hipotezą alternatywną ($\mu \neq \mu_0$) mamy

$$p(z) = P(Z \leqslant -|z| \text{ lub } Z \geqslant |z| \,|H_0) = 2P(Z \geqslant |z| \,|H_0) = 2(1 - \Phi(|z|)).$$

Natomiast dla hipotezy jednostronnej $\mu > \mu_0$ mamy

$$p(z) = P(Z \geqslant z|H_0) = 1 - \Phi(z).$$

Uwagi

1. p-wartość pozwala uniezależnić się do pewnego stopnia od arbitralnego wyboru poziomu istotności.
2. Małe p-wartości wspierają przekonanie o fałszywości H_0 i prawdziwości H_1.
3. Konwencjonalnie jako próg odrzucenia H_0 przyjmuje się 0,05.
4. p-wartość **nie jest** prawdopodobieństwem prawdziwości H_0, ani fałszywości H_1.

Test t

Test dla wartości średniej z **nieznanym** odchyleniem standardowym nazywa się **testem** t albo **testem Studenta**. Analogicznie do konstrukcji przedziałów ufności jako statystykę testową bierzemy

$$T = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}}.$$

Jeśli rozkład wyjściowy jest rozkładem normalnym i spełniona jest

$$H_0 : \mu = \mu_0,$$

to $T \sim t_{n-1}$. W innym przypadku prawdziwy rozkład T przybliża się przez t_{n-1}.

Test t — c. d.

Wobec tego zbiór krytyczny na poziomie istotności α dla dwustronnej hipotezy alternatywnej

$$H_1 : \mu \neq \mu_0$$

ma postać

$$C = \{t : t \leqslant -t_{1-\frac{\alpha}{2},n-1} \text{ lub } t \geqslant t_{1-\frac{\alpha}{2},n-1}\},$$

natomiast p-wartość w tym wypadku wynosi

$$p(t) = 2P(T \geqslant |t| \,|H_0) = 2(1 - F_{n-1}(|t|)),$$

gdzie F_{n-1} jest dystrybuantą t_{n-1}.

Przykład

```r
x <- rnorm(20, mean = 3, sd = 2)
t.test(x, mu = 2)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 0.96355, df = 19, p-value = 0.3474
## alternative hypothesis: true mean is not equal to 2
## 95 percent confidence interval:
##  1.461853 3.456324
## sample estimates:
## mean of x
##  2.459089
```

Test t — H_1 jednostronna $\mu > \mu_0$

W tym przypadku wartościami T wspierającymi odrzucenie H_0 na rzecz H_1 są duże liczby dodatnie, zatem na poziomie istotności α właściwy jest zbiór krytyczny

$$C = [t_{1-\alpha,n-1}, +\infty).$$

Natomiast p-wartość wynosi

$$p(t) = P(T \geqslant t | H_0) = 1 - F_{n-1}(t).$$

Przykład

```r
t.test(x, mu = 2, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  x
## t = 0.96355, df = 19, p-value = 0.1737
## alternative hypothesis: true mean is greater than 2
## 95 percent confidence interval:
##  1.635231      Inf
## sample estimates:
## mean of x
##  2.459089
```

Test t — H_1 jednostronna $\mu < \mu_0$

W tym przypadku wartościami T wspierającymi odrzucenie H_0 na rzecz H_1 są duże co do wartości bezwzględnej liczby ujemne, zatem na poziomie istotności α właściwy jest zbiór krytyczny

$$C = (-\infty, t_{\alpha,n-1}] = (-\infty, -t_{1-\alpha,n-1}].$$

Natomiast p-wartość wynosi

$$p(t) = P(T \leqslant t | H_0) = F_{n-1}(t).$$

Test t dla dwu prób — równe wariancje

Analogicznie jak w przypadku konstrukcji przedziałów ufności używamy statystyki

$$T = \frac{(\overline{X} - \overline{Y}) - (m_1 - m_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

i dla dwustronnej hipotezy alternatywnej

$$m_1 \neq m_2$$

na poziomie istotności α otrzymujemy zbiór krytyczny

$$C = (-\infty, -t_{1-\frac{\alpha}{2}, n_1+n_2-2}] \cup [t_{1-\frac{\alpha}{2}, n_1+n_2-2}, +\infty),$$

natomiast p-wartość wynosi

$$p(t) = 2P(T \geqslant |t| \,|H_0) = 2(1 - F_{n_1+n_2-2}(|t|)).$$

Przykład

```r
y <- rnorm(30, mean = 2, sd = 2)
t.test(x, y, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x and y
## t = 0.58564, df = 48, p-value = 0.5609
## alternative hypothesis: true difference in means is not equ
## 95 percent confidence interval:
##  -0.8506326  1.5498192
## sample estimates:
## mean of x mean of y
##  2.459089  2.109495
```

Test t dla dwu sparowanych prób

Zakładamy równe liczności X_i i Y_i oraz wspólny rozkład normalny różnic $D_i = X_i - Y_i$. Wtedy statystyką testową jest

$$T = \frac{X_i - Y_i - (m_1 - m_2)}{\frac{S_D}{\sqrt{n}}} \sim t_{n-1}$$

i dalszy ciąg jest jak w przypadku testu dla pojedynczej próby.

Przykład

```r
z <- rnorm(20, mean = 1, sd = 5)
t.test(x, z, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  x and z
## t = 2.4341, df = 19, p-value = 0.02497
## alternative hypothesis: true mean difference is not equal t
## 95 percent confidence interval:
##  0.3339504 4.4327913
## sample estimates:
## mean difference
##        2.383371
```

Test F równości wariancji I

Tutaj $X_1, \dots, X_{n_!} \sim \mathcal{N}(m_1, \sigma_1)$ i $Y_1, \dots, Y_{n_2} \sim \mathcal{N}(m_2, \sigma_2)$. Przyjmujemy

$$H_0 : \sigma_1^2 = \sigma_2^2.$$

Wtedy

$$F = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

i dla dwustronnej hipotezy alternatywnej

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

otrzymujemy na poziomie istotności α zbiór krytyczny

$$C = \left(0, f_{\frac{\alpha}{2}, n_1-1, n_2-1}\right] \cup \left[f_{1-\frac{\alpha}{2}, n_1-1, n_2-1}, +\infty\right),$$

Test F równości wariancji II

a p-wartość wynosi

$$p(f) = \begin{cases} 2F_{n_1-1,n_2-1}(f) & \text{dla } f < f_{1/2,n_1-1,n_2-1}, \\ 2(1 - F_{n_1-1,n_2-1}(f)) & \text{dla } f \geqslant f_{1/2,n_1-1,n_2-1}, \end{cases}$$

gdzie F_{n_1-1,n_2-1} jest dystrybuantą rozkładu $F(n_1 - 1, n_2 - 1)$.

Podobnie dla jednostronnej hipotezy alternatywnej

$$H_1 : \sigma_1^2 > \sigma_2^2$$

zbiór krytyczny na poziomie istotności α ma postać

$$C = [f_{1-\alpha,n_1-1,n_2-1}, +\infty),$$

a p-wartość wynosi

$$p(f) = 1 - F_{n_1-1,n_2-1}(f).$$

Przykład

```r
var.test(x, y)
```

```
##
##  F test to compare two variances
##
## data:  x and y
## F = 1.1066, num df = 19, denom df = 29,
## p-value = 0.7871
## alternative hypothesis: true ratio of variances is not equa
## 95 percent confidence interval:
##  0.4959402 2.6579380
## sample estimates:
## ratio of variances
##           1.106578
```

Przykład

```r
var.test(z, x, alternative = "greater")
```

```
##
##  F test to compare two variances
##
## data:  z and x
## F = 3.1032, num df = 19, denom df = 19,
## p-value = 0.008754
## alternative hypothesis: true ratio of variances is greater
## 95 percent confidence interval:
##  1.431192      Inf
## sample estimates:
## ratio of variances
##           3.103184
```